

J. Clin. Chem. Clin. Biochem.  
Vol. 19, 1981, pp. 121–137

## A Multivariate Approach for the Biometric Comparison of Analytical Methods in Clinical Chemistry

By U. Feldmann, B. Schneider, H. Klinkers†

Medical School Hannover, Department of Biometrics and Medical Informatics and

R. Haeckel

Medical School Hannover, Institute for Clinical Chemistry

(Received July 8/October 27, 1980)

**Summary:** The structural relationship model is recommended for the comparison of analytical methods in clinical chemistry. This model is based on the partition of the observed measurements of the different methods in 2 hypothetical random variables: the "expected values", which represent the correct value of the analyte with no error of measurement, and the "error term" representing the measurement errors. It is assumed, that both these variables are normally distributed.

There exists a linear structural relation between different analytical methods for the same analyte, provided the correlation coefficient between each pair of the expected values of these methods is 1. This linear structural relationship is expressed by the mean values  $\mu_i$  and standard deviations  $\alpha_i$  of the expected values, whereas the standard deviations of the error terms determine the precision of the methods. As a measure of the precision the coefficient of determination  $R_i^2$  is recommended.

The model of structural relationship is an extension of the well known regression models and gives a more realistic approach to the comparison of 2 or more analytical methods. With 2 methods the standardized principle component should therefore replace the regression analysis. The slope of this principle component is identical with the ratio  $s_y/s_x$ . Statistical methods for the estimation and for tests of hypotheses of the parameters are derived and demonstrated with an example.

### *Multivariate Verfahren zum biometrischen Vergleich analytischer Methoden in der Klinischen Chemie*

**Zusammenfassung:** Zum Vergleich verschiedener analytischer Bestimmungsmethoden in der klinischen Chemie werden die Strukturrelationsmodelle empfohlen. Diese Modelle basieren auf einer Zerlegung der beobachteten Werte in zwei hypothetische Zufallsgrößen: die „erwarteten Werte“, die den korrekten Wert des Analyts repräsentieren, wenn dieser ohne Meßfehler gemessen werden könnte, und die „Fehlerterme“, die Schwankungen aufgrund von Geräte- oder Meßfehler repräsentieren. In den Modellen wird im allgemeinen angenommen, daß beide Zufallsvariablen normalverteilt sind.

Das Vorliegen einer linearen Strukturrelation ist bei der Annahme der Normalverteilung äquivalent mit der Annahme, daß die erwarteten Werte von je zwei Verfahren streng miteinander korreliert sind; d. h. daß der Korrelationskoeffizient zwischen den beiden erwarteten Werten 1 ist. Diese lineare Strukturrelation kann durch die Mittelwerte  $\mu_i$  und die Standardabweichungen  $\alpha_i$  der erwarteten Werte vollständig beschrieben werden. Die Standardabweichungen der Fehlerterme charakterisieren die Präzision der jeweiligen Methoden. Als ein quantitatives Maß für die Präzision wird der Bestimmtheitskoeffizient  $R_i^2$  empfohlen, der als  $1 - \frac{\lambda_i^2}{\sigma_i^2}$  definiert ist ( $\lambda_i^2$  ist die Varianz des Fehlerterms und  $\sigma_i^2$  die Gesamtvarianz der Meßgrößen des Verfahrens Nr. i).

Die Strukturrelationsmodelle bilden eine Erweiterung der bekannten Regressionsmodelle und geben eine realistischere Basis für den Vergleich von zwei oder mehr analytischen Meßverfahren. Bei einem Vergleich von zwei Meßverfahren ist die lineare Strukturrelation (unter Annahme einer Normalverteilung) aus den Meßwerten nicht eindeutig zu identifizieren. Es muß vielmehr noch eine zusätzliche Nebenbedingung eingeführt werden, die im allgemeinen darin besteht, daß für das Verhältnis der beiden Fehlervarianzen ein fester Wert vorgegeben wird. Besondere Bedeutung für die An-

wendung besitzt die sogenannte standardisierte Hauptkomponentenmethode, bei der angenommen wird, daß dieses Verhältnis der Varianzen gleich der Neigung der Strukturgeraden ist. Diese Neigung kann dann einfach durch den Quotienten der beiden Standardabweichungen  $s_y/s_x$  geschätzt werden. Dieses Verhältnis ist besser geeignet, den Zusammenhang zwischen zwei Meßgrößen zu beschreiben, die beide zufällig schwanken, als der übliche Regressionskoeffizient.

In der Arbeit werden statistische Methoden referiert und teilweise neuentwickelt, die das Schätzen von Strukturrelationen und Tests über Hypothesen solcher Relationen auch in komplexeren Situationen (allgemeine Zusammenhänge zwischen mehreren Veränderlichen) gestatten. Die Methoden werden an einem Beispiel demonstriert.

## 1. Introduction

In clinical chemistry the number of analytical methods for measuring the value of the same chemical substance is increasing. To compare these different methods statistical methods are necessary, which give objective information on the precision and comparability of the various measurements. Often the methods of statistical regression analysis are used to solve this problem. But these methods have the disadvantage that in a pairwise comparison one of the measurements must be considered to be free of random errors. The regression line can depend greatly on the choice of this measurement (as "independent variable"). Therefore the methods of regression analysis cannot be considered as adequate for the comparison of analytical measurement methods, which are in general all influenced by random errors (see l.c. (1–6)).

It seems that *Deming* (7) was one of the first to introduce a statistical approach for the comparison of analytical methods in chemistry, which was not based on the assumption of the regression model. This approach was derived by *Deming* with geometric considerations and based on the principle of minimizing the orthogonal distance between the observed measurements and the estimated straight line (orthogonal mean square regression, see l.c. (8)). It can be shown that this principle is a specific feature of more general statistical methods, called *structural relationships*. These methods were originally introduced by *Karl Pearson* (9) and later discussed in different connections by several authors (see e.g. l.c. (10–18)). This approach is the adequate statistical way to compare two or more analytical values for the same analyte, where all measurements are influenced by random errors.

In this paper we will give an introduction to this approach and show its important role in clinical chemistry. The structural relationship will be demonstrated by comparing different analytical methods for the measurement of serum cholesterol. The relation to the regression analysis will be shown. Maximum likelihood estimates for the parameters (which characterize the dependence and precision of the different measures) and tests of hypothesis on these parameters (with or without constraints) are derived. The methods can be immediately extended to more complex relationships between the measurements or a more complex structure of the

random errors (e.g. hierarchical or factorial random effect models). They include the methods of factor analysis and path coefficient analysis (see l.c. (10, 19–24, 17, 5)).

## 2. The structural relationship approach for the comparison of 2 methods

### 2.1 The model

Let us assume that we have  $n$  samples of serum, drawn at random and independently from a population. In each of these samples an analyte is measured by two different methods, giving the measured values  $x_i$  and  $y_i$  for the  $i$ -th sample. These measurements are realizations of a pair of random variables ( $X, Y$ ) representing all possible pairs of values which can be measured in the population of all possible samples and measurements. As a basic model assumption the variation of these values is considered as additively composed of two components:

- the variation of the analyte within the population of all possible serum samples,
- the variation of the results in repeated measurements of the same serum sample.

The second variation is due to analytical errors. If such errors did not exist, each repetition of the measurement of the same sample would result in identical values. We call these values the "expected values" of the analyte in the sample and denote it by  $x_i^*$  and  $y_i^*$  for the  $i$ -th sample. The residuals between the measured sample values  $x_i$  and  $y_i$  and the corresponding expected values  $x_i^*$  and  $y_i^*$  indicate the analytical errors and are denoted by  $d_i$  and  $e_i$ . So we have the partition of the observed measurements in "expected values" and "error terms":

$$x_i = x_i^* + d_i$$

$$y_i = y_i^* + e_i$$

The expected values  $x_i^*$  and  $y_i^*$  vary within the population of all possible samples. So they can be considered as realizations of random variables  $X^*$  and  $Y^*$ , which represent the variation of the expected values within this population. We denote the mean values of these random variables by  $\mu_x$  and  $\mu_y$  and their variances by  $\alpha_x^2$  and  $\alpha_y^2$ . These variances indicate the variation of the expected values within the population of all possible samples and are therefore characteristics of this biological population.

The error terms  $d_i$  and  $e_i$  vary within the population of repeated measurements of the same sample. We assume that this variation is the same for all samples and can be represented by random variables  $D$  and  $E$ , from which the  $d_i$  and  $e_i$  are independent realizations. We assume that these random variables have mean values 0 and variances  $\lambda_x^2$  and  $\lambda_y^2$  resp. Furthermore, we assume that the error terms are mutually independent, independent from the random variables of the expected values and independent for different samples.

From these assumptions we conclude that the observed measurements  $x_i$  and  $y_i$  are realizations of two random variables  $X$  and  $Y$  with the mean values  $\mu_x$  and  $\mu_y$  (as the expected values) and the variances:

$$\sigma_x^2 = \alpha_x^2 + \lambda_x^2; \quad \sigma_y^2 = \alpha_y^2 + \lambda_y^2.$$

The expected values  $x_i^*$  and  $y_i^*$  indicate the correctness of the methods (if they agree with the proposed correct values for the sample), and the error terms  $d_i$  and  $e_i$  the precision. An adequate relative measure of the precision is the *coefficient of determination*  $R_x^2$  or  $R_y^2$ , which presents that part of the total variance  $\sigma_x^2$  or  $\sigma_y^2$ , which is determined by the variation of the expected values within the population of all possible samples, i.e.:

$$R_x^2 = \frac{\alpha_x^2}{\sigma_x^2} = 1 - \frac{\lambda_x^2}{\sigma_x^2}$$

$$R_y^2 = \frac{\alpha_y^2}{\sigma_y^2} = 1 - \frac{\lambda_y^2}{\sigma_y^2}.$$

If there are no measurement errors, the corresponding coefficient of determination is 1. In this case the corresponding error variances  $\lambda_x^2$  or  $\lambda_y^2$  are 0. If there is no variation of the expected values ( $\alpha_x^2$  or  $\alpha_y^2$  equal zero), all the variation of the observed measurements is due to measurement errors. Therefore the coefficient of determination is 0 in this case.

As both the analytical methods measure the same sample, there must be a relation between the expected values of the methods. We assume that this relation is a linear one and holds for all samples of the population. So we have:

$$y_i^* = \beta_0 + \beta_1 x_i^* \quad \text{for all values of } i,$$

where  $\beta_0$  is the intercept and  $\beta_1$  the slope (to the x-axis) of the corresponding straight-line.

As this linear relation holds for all possible samples, it must hold with probability one for the random variables:

$$Y^* = \beta_0 + \beta_1 X^* \quad \text{with probability one.}$$

This is the *structural relationship* equation for the two measurement methods. It combines the expected values of both methods and allows the prediction of the value of one measurement by knowing the value of the other one in a sample. The accuracy of this prediction depends on

the precision of the methods, which can be expressed by the coefficient of determination as defined above.

To get estimates for the two parameters  $\beta_0$  and  $\beta_1$  and a further insight in the structural relationship, additional assumptions on the probability distributions of the expected values and error terms are necessary. As usual in statistics, we assume that both the expected values  $X^*$ ,  $Y^*$  and the error terms  $D$ ,  $E$  have normal distributions with the parameters defined above. Then the random variables  $X$ ,  $Y$  of the observed measurements have a bivariate normal distribution with the mean values  $\mu_x$ ,  $\mu_y$ , the variances  $\sigma_x^2$ ,  $\sigma_y^2$  and a correlation  $\rho_{xy}$  determined by the structural relationship equation. In the x-y-plane the observed measurement values ( $x_i$ ,  $y_i$ ) are scattered like ellipses, as shown in figure 1 for 2 methods for the determination of the cholesterol concentration (obtained from l.c. (25)). These ellipses are called "dispersion ellipses" or "probability ellipses". The borderlines of these ellipses are the curves of constant probability density.

The structural relationship between the expected values  $X^*$  and  $Y^*$  implies that the bivariate normal distribution of these random values is degenerate, all the statistical mass being concentrated on a straight line, which of course must be the structural relationship-line (see l.c. (8)).

The implicit equation of this line is given by:

$$\frac{X^* - \mu_x}{\alpha_x} = \pm \frac{Y^* - \mu_y}{\alpha_y}$$

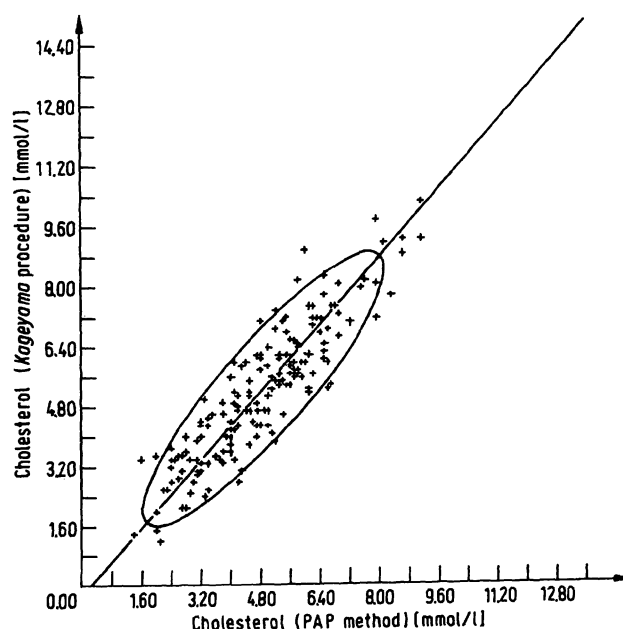


Fig. 1. Scattergram of the observed points ( $x_i$ ,  $y_i$ ) with principal component and dispersion ellipse. The data are taken from l.c. (14). PAP means a commercially available *Trinder*-method for determining the cholesterol concentration in serum samples.

where  $\mu_x$  and  $\mu_y$  are the mean values and  $\alpha_x$ ,  $\alpha_y$  the standard deviations of the expected values in the population of all samples. This shows that the intercept  $\beta_0$  and slope  $\beta_1$  of the structural relationship equation can be expressed uniquely by the mean values and standard deviations of the expected values:

$$\beta_0 = \mu_y \mp \frac{\alpha_y}{\alpha_x} \cdot \mu_x; \quad \beta_1 = \pm \frac{\alpha_y}{\alpha_x}.$$

This means that the structural relationship equation goes through the point of inertia ( $\mu_x$ ,  $\mu_y$ ) with a slope (except for the sign) given by the ratio of the two standard deviations of the expected values (i.e.  $\frac{\alpha_y}{\alpha_x}$  for the slope to the x-axis and  $\frac{\alpha_x}{\alpha_y}$  for that to the y-axis). As the structural relation equation is unique, it can be inverted to predict  $x_i^*$  from given  $y_i^*$ :

$$x_i^* = \beta'_0 + \beta'_1 y_i^*$$

with

$$\beta'_0 = \mu_x \mp \frac{\alpha_x}{\alpha_y} \cdot \mu_y, \quad \beta'_1 = \pm \frac{\alpha_x}{\alpha_y} = \frac{1}{\beta_1}.$$

Due to the linear structural relation between the expected values, the random variables of the observed values X and Y are correlated. Their covariance is given (except for the sign) by the product of the two standard deviations  $\alpha_x$  and  $\alpha_y$  of the expected values<sup>1)</sup>; the correlation coefficient is:

$$\rho_{x,y} = \pm \frac{\alpha_x \cdot \alpha_y}{\sigma_x \cdot \sigma_y}$$

(where the sign is determined by the sign of the real correlation between the two random variables).

So we see, that the following relations hold between the second-order central moments of the random variables X, Y and the structural relationship parameters:

$$\begin{aligned} \sigma_x^2 &= \alpha_x^2 + \lambda_x^2 & \sigma_y^2 &= \alpha_y^2 + \lambda_y^2 \\ \sigma_x^2 \cdot \sigma_y^2 \cdot \rho_{xy}^2 &= \alpha_x^2 \cdot \alpha_y^2. \end{aligned}$$

<sup>1)</sup> This follows from:

$$\text{cov}(X, Y) =$$

$$\epsilon(X - \mu_x)(Y - \mu_y) = \epsilon(X^* - \mu_x + E)(Y^* - \mu_y + D) =$$

$$= \epsilon[(X^* - \mu_x + E) \left( \pm \frac{\alpha_y}{\alpha_x} (X^* - \mu_x) + D \right)] = \pm \frac{\alpha_y}{\alpha_x} \epsilon(X^* - \mu_x)^2 \pm$$

$$\pm \frac{\alpha_y}{\alpha_x} \epsilon[E(X^* - \mu_x)] + \epsilon[D(X^* - \mu_x)] + \epsilon(ED) = \pm \alpha_y \alpha_x$$

(since E, D,  $X^*$  are assumed as mutually uncorrelated). The operator  $\epsilon$  means "expectation" (i.e.  $\epsilon(X) = \int x dF(x)$ ).

## 2.2 The problem of unidentifiability and its solutions

These relations show an intrinsic difficulty of the structural relationship approach for 2 measurements assuming normal distribution: There are only 5 estimable characteristics of the observable measurements X and Y ( $\mu_x$ ,  $\mu_y$ ,  $\sigma_x^2$ ,  $\sigma_y^2$  and  $\rho_{xy}^2$ ), but 6 unknown structural relationship parameters ( $\mu_x$ ,  $\mu_y$ ,  $\alpha_x^2$ ,  $\alpha_y^2$ ,  $\lambda_x^2$  and  $\lambda_y^2$ ).

So in general the structural relationship parameters cannot be identified by the observable pairs of measurements.

To overcome this difficulty several approaches are possible:

### a) Assumption of nonnormality

The assumption of a bivariate normal distribution for both measurement variables is replaced by other assumptions. There are two approaches discussed in the literature:

#### — Estimation via cumulants

This approach was introduced by Geary (13) and Reiersøl (17). It is based on the observation that the structural relationship involves linear relations between the cumulants of higher order (higher than 3) of the bivariate distribution. These cumulants or semi-invariants are characteristics of the distribution function uniquely related to the moments of the distribution (see l.c. (8), chapter 15.10). They can be estimated by the appropriate sample moments.

By these relations the slope  $\beta_1$  of the structural relationship line can be expressed as a function of the higher cumulants and therefore estimated by inserting in this function the estimated cumulants (or moments). The problem with this approach is that there is no unique relation between  $\beta_1$  and the cumulants; choosing different cumulants (of order higher than 3) results in different estimates for  $\beta_1$ . Geary suggests that cumulants of lowest order be used because of ease of computation. This would impose the rather strict assumption that the higher order cumulants are zero, which would be an objection to many practical situations. Other authors try to pool estimates of  $\beta_1$  based on different cumulants to get estimates of minimum variance. But this minimum depends on the exact form of the distribution function and may fail if the assumptions underlying the method are not correct. For details the interested reader is referred to l.c. (14).

#### — Estimating by grouping

This method in simplest form consists of ordering the observed pairs ( $x_i$ ,  $y_i$ ) in some manner, selecting proportions  $p_1$  and  $p_2$  such that  $p_1 + p_2 \leq 1$  (e.g.  $p_1 = p_2 = 1/3$ ), placing the first  $np_1$  pairs in one group  $G_1$  and the last  $np_2$  pairs in another group  $G_2$  (discarding the middle

group of observations if  $p_1 + p_2 < 1$ ) and estimating the slope  $\beta_1$  of the structural relationship line by the quotient of the weighted group-mean-differences of the  $y_i$  and  $x_i$  values; i.e. by:

$$b_1 = \left( \frac{1}{p_1} \sum_{i \in G_1} y_i - \frac{1}{p_2} \sum_{i \in G_2} y_i \right) / \left( \frac{1}{p_1} \sum_{i \in G_1} x_i - \frac{1}{p_2} \sum_{i \in G_2} x_i \right)$$

(if  $p_1 = p_2$ ; e.g. both are 1/3, this formula reduces to the ratio of the differences between the group-sums of the  $y_i$  and  $x_i$  values).

The problem is to choose an ordering manner which yields consistent and optimal estimation. This problem was treated by Wald (26), Neyman & Scott (27), and Wolfowitz (28, 18). A simple way would be to order the pairs  $(x_i, y_i)$  according to the rank of the  $x_i$ -values. But then rather complicated and unrealistic assumptions about the distributions of the pairs  $(x_i, y_i)$  must be made, if the estimate  $b$  is to be a consistent estimate (see Madansky (14)). Also other ordering rules lead to rather complicated conditions for the distribution of the measurement pairs, which are hard to realize and practically impossible to improve.

So as a summary of the discussion on the nonnormal approaches we cite Moran (15), p. 235: "These results are theoretically very complicated and not very useful in practice since it is only in very rare situations that we can be sure that the underlying distribution . . . is non-normal. It is also natural to expect that the closer this distribution approaches normality the more ineffective any such method of estimation will become." We will not further discuss this approach and refer the interested reader to the cited literature.

#### b) Repeated measurements

One gets independent estimates for  $\lambda_x^2$  and  $\lambda_y^2$  by repeating the measurements on the same samples  $p$ -times (e.g. twice or three times). So one gets pairs of measurements  $(x_{ij}, y_{ij})$ , where the first index is the number of the sample ( $i = 1, \dots, n$ ) and the second one the number of repetition ( $j = 1, \dots, p$ ). If the repeated measurements can be considered as realizations of independent random variables, having for each sample the same variance  $\lambda_x^2$  or  $\lambda_y^2$ , estimates for these variances are:

$$l_x^2 = \frac{1}{n(p-1)} \sum_{ij} (x_{ij} - \bar{x}_{i.})^2$$

$$l_y^2 = \frac{1}{n(p-1)} \sum_{ij} (y_{ij} - \bar{y}_{i.})^2$$

where  $\bar{x}_{i.} = \frac{1}{p} \sum_j x_{ij}$  and  $\bar{y}_{i.} = \frac{1}{p} \sum_j y_{ij}$  are the arithmetic means for the  $i$ -th sample.

If the measurements are repeated twice (double measurements) these formulas reduce to:

$$l_x^2 = \frac{\sum_i (x_{i1} - x_{i2})^2}{2n}$$

$$l_y^2 = \frac{\sum_i (y_{i1} - y_{i2})^2}{2n}$$

Estimates for the total variances  $\sigma_x^2$  and  $\sigma_y^2$  are:

$$s_x^2 = \frac{1}{np-1} \sum_{ij} (x_{ij} - \bar{x}_{..})^2 \quad s_y^2 = \frac{1}{np-1} \sum_{ij} (y_{ij} - \bar{y}_{..})^2$$

where  $\bar{x}_{..} = \frac{1}{np} \sum_{ij} x_{ij}$  and  $\bar{y}_{..} = \frac{1}{np} \sum_{ij} y_{ij}$  are the overall means.

From this one gets as estimates for the variances  $\alpha_x^2$  and  $\alpha_y^2$ :

$$a_x^2 = s_x^2 - l_x^2 \quad a_y^2 = s_y^2 - l_y^2$$

and for the slope  $\beta_1$  of the structural relationship equation:

$$b_1 = \frac{a_y}{a_x} = \pm \frac{\sqrt{s_y^2 - l_y^2}}{\sqrt{s_x^2 - l_x^2}}$$

(where the sign is the same as for the estimated covariance between X and Y).

In general these estimates  $a_x^2$  and  $a_y^2$  will not fit the third relation, which reads in estimates:

$$s_x^2 \cdot s_y^2 \cdot r_{xy}^2 = a_x^2 \cdot a_y^2$$

where  $r_{xy}$  is the sample correlation of X and Y.

As the analytical errors of the measurements within an analyte sample are supposed as uncorrelated, this sample correlation is given by the correlation between the sample means of the two measurements within the repeated measurements of the analyte samples:

$$r_{xy} = \frac{1}{n-1} \sum_i (x_{i.} - \bar{x}_{..})(y_{i.} - \bar{y}_{..}) / s_x s_y$$

Using this sample correlation and the relation ahead one gets as estimate of the slope  $\beta_1$ :

$$\tilde{b}_1 = \frac{s_y^2 - l_y^2}{s_x s_y r_{xy}}$$

This estimate is in general different from the first one; both are consistent (i.e. they converge for increasing  $n$  to the true parameter  $\beta_1$ ) but not efficient estimates (estimates with at least asymptotic smallest variance).

Madansky (14) has shown, that the whole problem can be treated as a variance component problem, where the total variance and covariance of the two measurements is

decomposed in the variation between the analyte samples and within the repeated measurements of the sample. These variance components generate additional consistent, but not efficient estimates for  $\beta_1$ .

The problem is to make an appropriate choice of the "best" of these consistent estimates. This problem, which is known as "overidentification" problem was discussed by Barnett (11) and Kiefer (29), who showed that an efficient estimate can be obtained by introducing the estimates  $l_x^2$  and  $l_y^2$  (together with the estimates  $s_x^2$ ,  $s_y^2$  and  $r_{xy}$  derived above) into the maximum likelihood equation and solving it with respect to  $\beta_1$ . This leads to the estimate  $b_1$  given in table 1 for the constraint  $\Theta^2 = l_y^2/l_x^2$ .

It should be mentioned that a unique consistent and efficient estimate of  $\beta_1$  is possible, if only one of the measurements is repeated. If e.g. only the  $x_{ij}$ -measurements are repeated, one can directly estimate  $l_x^2$  and get an efficient estimate of  $\beta_1$ :

$$b_1 = \frac{s_x s_y r_{xy}}{s_x^2 - l_x^2}$$

If the y-measurements are repeated ( $y_{ij}$ ) the residual variance  $l_y^2$  can be directly estimated and the estimate of  $\beta_1$  is:

$$b_1 = \frac{s_y^2 - l_y^2}{s_x s_y r_{xy}}$$

The estimates of the sample correlation  $r_{xy}$  are:

$$r_{xy} = \frac{1}{n-1} \sum_i (x_i - \bar{x})(y_i - \bar{y})/s_x s_y$$

if measurement x is repeated

$$r_{xy} = \frac{1}{n-1} \sum_i (x_i - \bar{x})(y_i - \bar{y})/s_x s_y$$

if measurement y is repeated.

### c) Constraints on the error-variances

The approach mostly used to overcome the unidentifiability problem is to impose constraints on the structural relationship parameters. There are several types of such constraints possible, two of which are of particular importance:

The *first type* of constraints assumes either  $\lambda_x$  or  $\lambda_y$  equal to 0. This means, that either the first or the second measurement method is free of measurement errors. It is analogous to the assumptions in regression analysis and reduces the structural relationship model to the corresponding regression models:

$$y_i = \beta_0 + \beta_1 x_i + e_i \quad \text{if } \lambda_x = 0$$

$$x_i = \beta'_0 + \beta'_1 y_i + d_i \quad \text{if } \lambda_y = 0.$$

But there is an important difference between these models and the usual regression analysis. Whereas in the

Tab. 1. The slope  $b_1$  and the standard deviations of random errors  $l_x$ ,  $l_y$  are given in terms of the standard deviations  $s_x$ ,  $s_y$  and the correlation  $r$  of the two observed measurements.

The constant  $c_1$  in principle component analysis is given by:  $c_1 = +\sqrt{(s_y^2 - s_x^2)^2 + 4r^2 \cdot s_x^2 \cdot s_y^2}$

The constant  $c_2$  in the model with constraint  $\lambda_y^2/\lambda_x^2 = \Theta^2$  is given by:  $c_2 = \sqrt{(s_y^2 - \Theta^2 s_x^2)^2 + 4\Theta^2 r^2 s_x^2 s_y^2}$

Statistical analysis	Slope of linear relationship $b_1$	Standard deviation of random error	
		Measurement x $l_x$	Measurement y $l_y$
Univariate	Regression (x independent) $\frac{r \cdot s_y}{s_x}$	0	$\sqrt{\frac{n-1}{n-2} s_y^2 (1-r^2)}$
	Regression (y independent) $\frac{s_y}{r \cdot s_x}$	$\sqrt{\frac{n-1}{n-2} \cdot s_x^2 (1-r^2)}$	0
Bivariate	Principle component $\frac{1}{2 \cdot r \cdot s_x \cdot s_y} (s_y^2 - s_x^2 + c_1)$	$\sqrt{\frac{1}{2} \cdot \frac{n}{n-1} (s_y^2 + s_x^2 - c_1)}$	$\sqrt{\frac{1}{2} \cdot \frac{n}{n-1} (s_y^2 + s_x^2 - c_1)}$
	Standardized principle component $\frac{r}{ r } \cdot \frac{s_y}{s_x}$	$\sqrt{\frac{n}{n-1} s_x^2 \cdot (1- r )}$	$\sqrt{\frac{n}{n-1} s_y^2 \cdot (1- r )}$
Constraint $\lambda_y^2/\lambda_x^2 = \Theta^2$	$\frac{1}{2 \cdot r \cdot s_x \cdot s_y} (s_y^2 - \Theta^2 s_x^2 + c_2)$	$\sqrt{\frac{1}{2\Theta^2} \cdot \frac{n}{n-1} (s_y^2 + \Theta^2 s_x^2 - c_2)}$	$\sqrt{\frac{1}{2} \cdot \frac{n}{n-1} (s_y^2 + \Theta^2 s_x^2 - c_2)}$

regression analysis usually the  $x_i$  or  $y_i$  are considered as given fixed values, in the structural relationship approach these values are realizations of random variables  $X$  or  $Y$  with the variance  $\sigma_x^2 = \alpha_x^2$  or  $\sigma_y^2 = \alpha_y^2$ .

The regression coefficients  $\beta_1$  or  $\beta'_1$  are related to the second order moments of the measurements  $X$  or  $Y$  according to the following equations:

$$\beta_1 = \frac{\sigma_y}{\sigma_x} \rho_{xy} \quad \text{if } \lambda_x = 0$$

$$\beta'_1 = \frac{\sigma_y}{\sigma_x} \rho_{xy} \quad \text{if } \lambda_y = 0.$$

For the remaining error variances we get:

$$\lambda_y^2 = \sigma_y^2(1 - \rho_{xy}^2) \quad \text{if } \lambda_x = 0$$

$$\lambda_x^2 = \sigma_x^2(1 - \rho_{xy}^2) \quad \text{if } \lambda_y = 0.$$

The *second type* of constraints assumes that the ratio of the error variances takes a fixed value  $\Theta^2$ :

$$\frac{\lambda_y^2}{\lambda_x^2} = \Theta^2.$$

The value  $\Theta^2$  is either assumed as known or related to other structural relationship parameters.

With this type of constraint the number of structural relationship parameters is reduced to 3 independent parameters ( $\alpha_x^2$ ,  $\alpha_y^2$  and  $\lambda_x^2$ , whilst  $\lambda_y^2$  depends on  $\lambda_x^2$  by the relation  $\lambda_y^2 = \Theta^2 \lambda_x^2$ ), which can be related uniquely to the second order moments ( $\sigma_x^2$ ,  $\sigma_y^2$  and  $\rho_{xy}$ ) of the measurements. So we get:

$$\alpha_x^2 = -\frac{\sigma_y^2 - \Theta^2 \sigma_x^2}{2\Theta^2} + \frac{1}{2\Theta^2} \sqrt{(\sigma_y^2 - \Theta^2 \sigma_x^2)^2 + 4\Theta^2 \rho_{xy}^2 \sigma_x^2 \sigma_y^2}$$

$$\alpha_y^2 = \frac{\sigma_y^2 - \Theta^2 \sigma_x^2}{2} + \frac{1}{2} \sqrt{(\sigma_y^2 - \Theta^2 \sigma_x^2)^2 + 4\Theta^2 \rho_{xy}^2 \sigma_x^2 \sigma_y^2}$$

$$\lambda_x^2 = \frac{1}{2\Theta^2} (\sigma_y^2 + \Theta^2 \sigma_x^2 - \sqrt{(\sigma_y^2 - \Theta^2 \sigma_x^2)^2 + 4\Theta^2 \rho_{xy}^2 \sigma_x^2 \sigma_y^2}).$$

From these relations we get, for the slope  $\beta_1$  of the structural relationship line to the x-axis, the expression:

$$\beta_1 = \pm \frac{\alpha_y}{\alpha_x} = \frac{(\sigma_y^2 - \Theta^2 \sigma_x^2) + \sqrt{(\sigma_y^2 - \Theta^2 \sigma_x^2)^2 + 4\Theta^2 \rho_{xy}^2 \sigma_x^2 \sigma_y^2}}{2\rho_{xy} \sigma_x \sigma_y}.$$

Two assumptions about  $\Theta^2$  have particular interest:

*First assumption* (principal component model):

$$\Theta^2 = 1 \quad \text{or} \quad \lambda_x^2 = \lambda_y^2.$$

The error variances of both methods are considered as equal. One can show mathematically (see e.g. l.c. (8)) that in this case the structural relationship line is identical with the major axis of the dispersion ellipses (formed by the elements of the covariance matrix) for the bivariate normal distribution of the measurement variables  $X$  and  $Y$ . This major axis is also called the *principal component*. Its slope  $\beta_1$  to the x-axis is given by:

$$\beta_1 = \frac{(\sigma_y^2 - \sigma_x^2) + \sqrt{(\sigma_y^2 - \sigma_x^2)^2 + 4\rho_{xy}^2 \sigma_x^2 \sigma_y^2}}{2\rho_{xy} \sigma_x \cdot \sigma_y}$$

The residual variance  $\lambda^2$  (which in this case is the same for D and E) is:

$$\lambda^2 = \frac{1}{2} (\sigma_x^2 + \sigma_y^2) - \frac{1}{2} \sqrt{(\sigma_y^2 - \sigma_x^2)^2 + 4\rho_{xy}^2 \sigma_x^2 \sigma_y^2}.$$

We call the model with this assumption the "principal component model".

*Second assumption* (standardized principal component model):

$$\Theta^2 = \beta_1^2 \quad \text{or} \quad \frac{\lambda_x^2}{\sigma_x^2} = \frac{\lambda_y^2}{\sigma_y^2}.$$

By this assumption the ratio of the analytical error variance to the total variance is equal for both measurement methods. This is equivalent to the assumption of equal error variances for the standardized measurement values  $(X - \mu_x)/\sigma_x$  and  $(Y - \mu_y)/\sigma_y$ . Therefore the structural relationship line corresponds in this case with the principal component of the standardized measurements, i.e. the major axis of the correlation ellipses (the ellipses defined by the quadratic form with the elements of the correlation matrix as coefficients). We call this model the "standardized principal component" model. The slope  $\beta_1$  of the structural relationship line to the x-axis is given by:

$$\beta_1 = \pm \frac{\sigma_y}{\sigma_x}$$

where the sign is equal to the sign of the correlation coefficient  $\rho_{xy}$ .

An interesting point to mention is the fact that the slope of the standardized principal component model is the geometric mean of the slopes in the two regression models. This property was used by *Averdunk & Borner* (1) to define a unique regression line between two measurements  $x_i$  and  $y_i$  (see also l.c. (15)).

The error variances  $\lambda_x^2$  and  $\lambda_y^2$  are given by:

$$\lambda_x^2 = \sigma_x^2(1 - |\rho_{xy}|) \quad \lambda_y^2 = \sigma_y^2(1 - |\rho_{xy}|).$$

The coefficient of determination equals for both measurement variables the absolute value of the correlation coefficient. Comparing this with the coefficient of determination in the case of the regression model, which equals the square of the correlation coefficient, we see that in the standardized principal component model the coefficient of determination is mostly greater (and only for  $\rho = \pm 1$  and  $\rho = 0$  equal) than that of the regression model.

It should be mentioned that the constraint  $\lambda_y^2/\lambda_x^2 = \Theta^2$  can be reduced to the principal component model by recoding the  $x_i$  variables in  $\tilde{x}_i = \Theta x_i$ . The structural relationship line between  $y_i$  and  $\tilde{x}_i$  has the slope  $\tilde{\beta}_1 = \beta_1/\Theta$ .

As all the structural relationship lines go through the point of inertia ( $\mu_x, \mu_y$ ) of the measurement pairs, the intercept  $\beta_0$  of the straight line is for all possible constraints given by:

$$\beta_0 = \mu_y - \beta_1 \mu_x.$$

Concerning the interrelation of the structural relationship lines with different constraints, it is important to know that all lines with finite positive ratios of error variances (i.e. with  $0 < \lambda_y^2/\lambda_x^2 = \Theta^2 < \infty$ ) lay between the two regression lines:  $Y^* = \beta_0 + \beta_1 X$  and  $X^* = \beta'_0 + \beta'_1 Y$  (see fig. 3). These two lines correspond to the limiting values of  $\Theta^2 \rightarrow 0$  and  $\Theta^2 \rightarrow \infty$ . The two regression lines are rather close together. Thus if the square of the correlation is high (e.g. higher than 0.9), the choice of the constraint is rather unimportant, but it may be very important in the case of low correlation between the two measurement variables.

### 2.3 Statistical estimation and test of significance

If the unidentifiability problem is solved by using constraints of the form  $\frac{\lambda_y^2}{\lambda_x^2} = \Theta^2$ , the structural relationship

parameters are uniquely related to the second order moments  $\sigma_x^2$ ,  $\sigma_y^2$  and  $\sigma_{xy}$ . Therefore they can be optimally estimated by inserting the maximum likelihood estimates of the moments into the relations between moments and parameters. The maximum likelihood estimates of moments are:

$$S_x^2 = \frac{1}{n} \sum_i (x_i - \bar{x})^2 = \frac{n-1}{n} s_x^2$$

$$S_y^2 = \frac{1}{n} \sum_i (y_i - \bar{y})^2 = \frac{n-1}{n} s_y^2$$

$$S_{xy} = \frac{1}{n} \sum_i (x_i - \bar{x})(y_i - \bar{y}) = \frac{n-1}{n} s_{xy}$$

where  $s_x^2$ ,  $s_y^2$  and  $s_{xy}$  are the usual unbiased estimates of the variances  $\sigma_x^2$ ,  $\sigma_y^2$  and the covariance  $\sigma_{xy}$ . The estimate of the correlation coefficient  $\rho_{xy}$  is:

$$r_{xy} = \frac{S_{xy}}{S_x \cdot S_y} = \frac{s_{xy}}{s_x \cdot s_y} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}.$$

Using these estimates, together with the relations derived ahead, the maximum likelihood estimates for the slope  $\beta_1$  of the structural relationship line and the standard deviations  $\lambda_x$  and  $\lambda_y$  of the residuals are expressed by the formulas shown in table 1.

The different lines resulting from the different constraints have an interesting geometrical interpretation. All lines minimize the mean square distance between the observed measurements ( $x_i, y_i$ ) — represented as points in a plane — and a straight line; but the distance being measured in different directions depends on the constraint  $\Theta$ . The slope of the direction of the distance to the x-axis is given by:  $\delta = -\Theta^2/\beta_1$ . For the regression line to the x-axis the direction of the distance is orthogonal to the x-axis or parallel to the y-axis (see fig. 2a). For the regression to the y-axis the distance is measured and minimized parallel to the x-axis (see fig. 2b). For the principal component model the distance is measured and minimized orthogonal to the straight line (see fig. 2c), which means that the slope of the distance to the x-axis is  $-1/\beta_1$ . This structural relationship line is therefore also called the "orthogonal regression line" (see e.g. l.c. (8, 9)). In the standardized principal component model the direction of the distance to be minimized has the slope  $-\beta_1$  to the x-axis. This means it forms, with the x-axis, the same angle as the estimated structural relationship line, but in the opposite direction (see fig. 2d).

In figure 2 these 4 different structural relationship lines and the directions of the distance measure are plotted into the scattergram of serum cholesterol measurements obtained by 2 different methods. The experimental data are taken from l.c. (25).

The estimates of the first and second order moments were calculated from these 165 pairs of measurements. They are summarized in table 2. Table 3 shows the corresponding estimates  $b_0$  and  $b_1$  for the intercept and the slope to the x-axis and the estimates  $l_x$  and  $l_y$  for the residual standard deviations derived under the different constraints. The scattergram of the observed points and the 4 structural relationship lines are shown in figure 3. As was pointed out before, all the lines are enclosed by the 2 regression lines. Since the correlation coefficient is near to 1 (0.893), the various structural relationship lines do not differ very much.

The sampling distribution of the slope estimate  $b_1$  is not known and even approximative values for the variance are complicated to derive. An approach to asymptotic



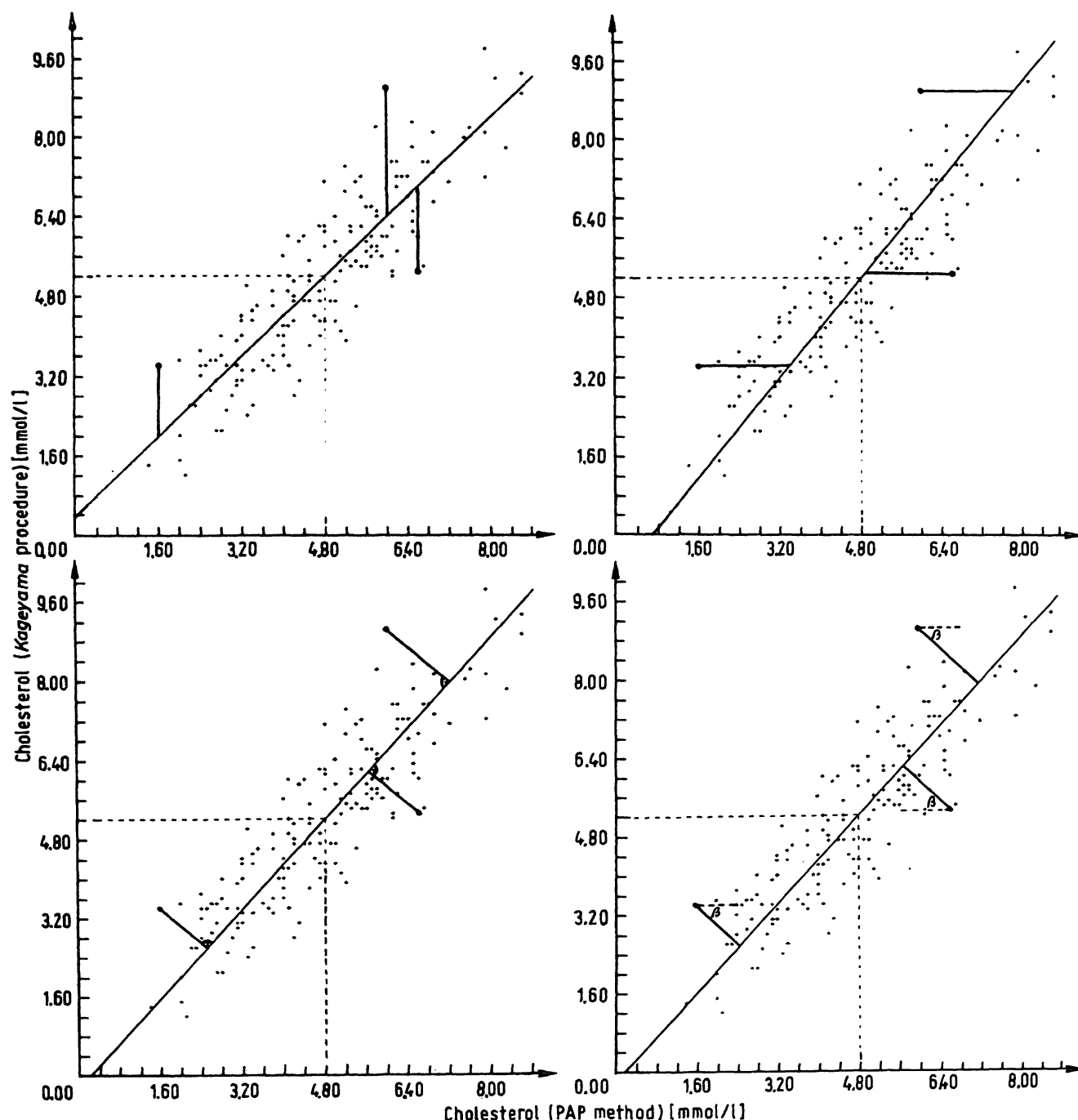


Fig. 2. The linear structural relationship with 4 different types of constraints

a) regression from y to x:  $\lambda_x = 0$  ( $\Theta = \infty$ )

c) principle component:  $\lambda_x^2 = \lambda_y^2$  ( $\Theta = 1$ )

b) regression from x to y:  $\lambda_y = 0$  ( $\Theta = 0$ )

d) standardized principle component:  $\frac{\lambda_x^2}{\sigma_x^2} = \frac{\lambda_y^2}{\sigma_y^2}$  ( $\Theta = \beta_1$ )

The bars show the direction of distances to be minimized.

expressions for the variances was given by *Barnett* (12), using the fact that the estimates of the structural relationship parameters are functions of the sample moments  $S_x^2$ ,  $S_y^2$  and  $S_{xy} = r \cdot S_x \cdot S_y$ . For such functions approximative expressions for the variance are known (see l.c. (8) and appendix 1). Using these expressions we get, for the estimate  $b_1 = \pm s_y/s_x$  in the standardized principal component model (which is most important in practice),

the following approximative expression for the variance (with an additive error proportional to  $(1/n)^{3/2}$ ):

$$s_{b_1}^2 = \frac{1}{n} b_1^2 (1 - r_{xy}^2).$$

The details of the derivation of this formula and approximative expressions for the slope under other constraints are given in appendix 1.

Tab. 2. Sample statistics for two methods determining the cholesterol concentration. The experimental data are taken from l.c. (14). Both methods use cholesterol-oxidase, "PAP" the *Trinder* reaction and „Kageyama“ the catalase coupled *Hantzsch* reaction for the  $H_2O_2$  detection.

Sample statistics	Methods for determining the cholesterol concentration [mmol/l]	
	X = „PAP“	Y = „Kageyama“
Means	$\bar{x} = 4.799$	$\bar{y} = 5.198$
Variances	$s_x^2 = 2.625$	$s_y^2 = 3.324$
Standard deviations	$s_x = 1.620$	$s_y = 1.823$
Covariance	$s_{xy} = 2.638$	
Correlation	$r = 0.893$	
No. of samples	$n = 165$	

Tab. 3. The estimated linear relationship  $Y^* = b_0 + b_1 X^*$  between X = „PAP“ and Y = „Kageyama“ calculated from table 1 with the sample statistics of table 2.

Statistical analysis	Intercept	Slope	Standard deviation of random error	
	$b_0$	$b_1$	$l_x$	$l_y$
Regression to x	0.377	1.005	0.000	0.824
Regression to y	-0.849	1.260	0.731	0.000
Principle component	-0.278	1.141	0.560	0.560
Standardized principle component	-0.201	1.125	0.530	0.596

In regression analysis (considering regression to the x-axis), it is usually stated that the slope  $b_1$  is normally distributed with mean value  $\beta_1$  and standard deviation  $\sigma_{\beta_1} = \sigma_e / \sqrt{\sum (x_i - \bar{x})^2}$ , where  $\sigma_e$  is the standard deviation of the residuals. (In the notation used in this paper we called it  $\lambda_y$ ). It should be noticed that this statement is true for small values of n, only if the values  $x_i$  are considered as fixed and not as realizations of random variables (as in the structural relationship model). So this distributional property of  $b_1$  in the regression model is conditional on the values  $x_i$ . Therefore the confidence intervals and tests derived under this distributional property are also conditional on the given  $x_i$ -values. Similar results hold for the regression slope  $b'_1$  to the y-axis, where the values  $y_i$  are considered as fixed.

Of the possible hypotheses to be tested, the hypothesis  $\beta_1 = 1$  is the most important one. It states that both measurement methods have the same expected values, what can be interpreted as lack of proportional errors in both measurement methods. It should be noted, that this hypothesis is an additional constraint. As the structural relationship parameters are uniquely related to the second order moments by the constraint  $\lambda_y^2 / \lambda_x^2 = \Theta^2$ , the

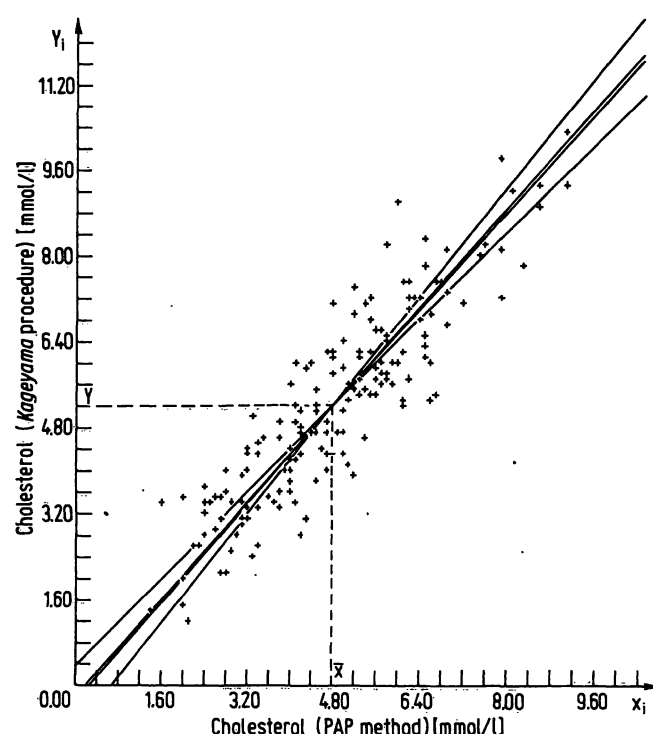


Fig. 3. The different types of linear structural relationship applied to the same data points as shown in figure 1.

hypothesis  $\beta_1 = 1$  imposes a constraint to the second order moments, which may be written as:

$$\sigma_y^2 = \Theta^2 \sigma_x^2 + \sigma_x \sigma_y \rho_{xy} (1 - \Theta^2) \quad (\text{for } \beta_1 = 1).$$

In the case of the principal component and standardized principal component model (which are identical for  $\beta_1 = 1 = \Theta$ ) this constraint reduces to:

$$\sigma_y^2 = \sigma_x^2.$$

So we see that for both these models the hypothesis  $\beta_1 = 1$  is equivalent to the hypothesis  $\sigma_y^2 = \sigma_x^2$  (where X and Y are dependent).

A test for this hypothesis was given by *Morgan* (30). This test is based on a linear transformation of the measurements  $x_i, y_i$  to  $u_i, v_i$  according to:

$$u_i = \frac{1}{2} (x_i + y_i); \quad v_i = \frac{1}{2} (x_i - y_i).$$

These transformed measurements are realizations of random variables U, V normally distributed with the parameters:

$$\mu_u = \frac{1}{2} (\mu_x + \mu_y); \quad \mu_v = \frac{1}{2} (\mu_x - \mu_y)$$

$$\sigma_u^2 = \frac{1}{4} (\sigma_x^2 + \sigma_y^2 + 2\sigma_{xy}); \quad \sigma_v^2 = \frac{1}{4} (\sigma_x^2 + \sigma_y^2 - 2\sigma_{xy})$$

$$\rho_{uv} = \frac{\sigma_x^2 - \sigma_y^2}{\sqrt{(\sigma_x^2 + \sigma_y^2)^2 - 4\sigma_x^2 \sigma_y^2 \rho_{xy}^2}}$$

where  $\mu_x, \mu_y, \sigma_x^2, \sigma_y^2$  and  $\rho_{xy}$  are the parameters of the bivariate normal distribution of the measurement pairs (X, Y).

From these relations it is obvious that the hypothesis  $\sigma_x^2 = \sigma_y^2$  is equivalent to the hypothesis  $\rho_{uv} = 0$ . This hypothesis can be tested using the sample correlation coefficient  $r_{uv}$ , i.e.

$$r_{uv} = \frac{s_x^2 - s_y^2}{\sqrt{(s_x^2 + s_y^2)^2 - 4s_x^2 s_y^2 r_{xy}^2}}$$

According to a well known result (first derived by Gosset writing under the pseudonym "Student" in 1908; see l.c. (8), chapter 29.7) the sample statistic

$$t = \frac{r_{uv} \sqrt{n-2}}{\sqrt{1-r_{uv}^2}}$$

is distributed as Student's t-statistic with  $n-2$  degrees of freedom under the null hypothesis  $\rho_{uv} = 0$ .

Using this result for the case of the principal component and standardized principal component model the hypothesis  $H_0: \beta_1 = 1$  can be tested by the statistic:

$$t = \sqrt{n-2} \frac{s_x^2 - s_y^2}{2s_x s_y \sqrt{1-r_{xy}^2}}$$

which is t-distributed under  $H_0$  with  $n-2$  degrees of freedom.

With a two-sided alternative hypothesis ( $H_1: \beta_1 \neq 1$ ) the null hypothesis  $H_0$  is to be rejected with an error probability  $\alpha$  (e.g.  $\alpha = 0.05$ ), if the absolute value  $|t|$  exceeds the  $(1 - \frac{\alpha}{2})$ -quantile of the t-distribution. These quantiles are for various degrees of freedom tabulated in nearly all statistical text books.

Applying this test to the two methods for serum cholesterol concentration (25) we get from table 2:

$$s_x^2 = 2.625; s_y^2 = 3.324; s_{xy} = 2.638; r_{xy}^2 = 0.798; n = 165.$$

From these estimates follows:

$$t = \sqrt{163} \frac{-0.699}{2.655} = -3.36.$$

With  $n-2 = 163$  degrees of freedom the 0.9995-quantile of the t-distribution (which corresponds to a two-sided error probability  $\alpha = 0.001$ ) is 3.35. As the absolute value of the calculated  $t$  (i.e. 3.36) exceeds this level, the null hypothesis  $\beta_1 = 1$  can be rejected with an error probability of less than 0.1%. The slope of the structural relationship line is very significantly different from 1, which indicates a proportional error of both measurements.

This test can be immediately extended to test general simple null hypotheses of the form  $H_0: \beta_1 = \beta_{10}$ , where  $\beta_{10}$  is any given value.

To test this hypothesis we rescale the  $x_i$ -values to  $\tilde{x}_i = \beta_{10} x_i$  (the  $y_i$  values remain unchanged). If  $H_0$  is correct, the structural relationship line between  $y_i$  and  $\tilde{x}_i$  has the slope  $\tilde{\beta}_1 = 1$ . So we have reduced the general null hypothesis  $\beta_1 = \beta_{10}$  to the specific one:  $\tilde{\beta}_1 = 1$  and can apply the test derived above.

It should be noticed that rescaling the  $x_i$  changes the variances and constraint  $\Theta$ . The error variances of the variables  $\tilde{x}_i$  and  $y_i$  have the constraint  $\tilde{\Theta}^2 = \lambda_y^2 / \lambda_x^2 = \Theta^2 / \beta_{10}^2$ . If the original measurements follow the standardized principal component model with the constraint (under  $H_0$ )  $\Theta^2 = \beta_{10}^2$ , the rescaled measurements  $\tilde{x}_i, y_i$  follow the principal component model with the constraint  $\tilde{\Theta}^2 = 1$ . For this model the hypothesis  $\beta_1 = \beta_{10}$  is equivalent to the hypothesis  $\tilde{\beta}_1 = 1$ .

The test of the general simple null hypothesis  $H_0: \beta_1 = \beta_{10}$  uses as test statistic:

$$t(\beta_{10}) = \sqrt{n-2} \frac{\beta_{10}^2 s_x^2 - s_y^2}{2\beta_{10} s_x s_y \sqrt{1-r_{xy}^2}}$$

This statistic is t-distributed under  $H_0$  with  $n-2$  degrees of freedom. If  $|t(\beta_{10})| > t_p$ , where  $P = 1 - \alpha/2$  and  $t_p$  is the P-quantile of the t-distribution with  $n-2$  degrees of freedom, the hypothesis  $H_0$  is to be rejected with a two-sided error probability of  $\alpha$ .

In this general test the statistic  $t$  is a function of the slope  $\beta_1$  assumed in the null hypothesis. Fixing the quantile  $t_p$  we can use this function to calculate confidence bounds for  $\beta_1$  to the given confidence probability  $P$ , as these bounds are defined by the equation:

$$t(\beta_1) = \sqrt{n-2} \frac{\beta_1^2 s_x^2 - s_y^2}{2\beta_1 s_x s_y \sqrt{1-r_{xy}^2}} = \pm t_p.$$

This is a quadratic equation in  $\beta_1$ . The solutions of this equation are

$$\beta_{1,u,l} = \pm \frac{s_y}{s_x} \left( \sqrt{1 + (1-r_{xy}^2) \frac{t_p^2}{n-2}} \pm \frac{t_p}{\sqrt{n-2}} \sqrt{1-r_{xy}^2} \right).$$

The first sign corresponds to the sign of  $\beta_1$  which equals the sign of the correlation coefficient  $r_{xy}$ . The second sign gives the upper and lower confidence limits. This confidence interval is consistent, i.e. for increasing  $n$  both confidence limits converge to the parameter  $\beta_1$ , i.e. the true slope of the structural relationship line in the standardized principal component model. By rearranging the terms in the brackets, we can write for this confidence interval:

$$\beta_{1,u,l} = b_1 + B \pm \sqrt{\frac{n}{n-2}} t_p s_{b_1}$$

where  $b_1 = \pm \frac{s_y}{s_x}$  is the estimate for the slope  $\beta_1$ ,  $B$

( $B = b_1(\sqrt{1 + (1 - r_{xy}^2) \frac{t_p^2}{n-2}} - 1)$ ) is the bias,

$$s_{b_1} = \frac{1}{\sqrt{n}} b_1 \sqrt{1 - r_{xy}^2}$$

is the (asymptotic) standard deviation of  $b_1$  and  $t_p$ , the P-quantile of the t-distribution with  $n - 2$  degrees of freedom.

For the example of the two cholesterol measurements we get (see tab. 2 and 3):

$$b_1 = 1.125; s_x^2 = 2.625; s_y^2 = 3.324; r_{xy}^2 = 0.798; n = 165.$$

Assuming  $P = 0.95$  we find  $t_p = 1.975$  we get:

$$\text{upper confidence bound: } \beta_{1u} = 1.206$$

$$\text{lower confidence bound: } \beta_{1l} = 1.050.$$

As the confidence interval does not cover the value 1, we conclude that the slope  $\beta_1$  is significantly different from 1 (as shown by the t-test).

### 3. Extension to the comparison of more than two methods

#### 3.1 Extension of the structural relationship model

The model of linear structural relationship can immediately be extended to the comparison of more than two analytic methods measuring the value of the same chemical substance (see l.c. (3, 23, 17, 5)). Let us assume that there are  $m$  such methods ( $m > 2$ ). The observed values of method number  $j$  in the serum sample number  $k$  will be denoted by  $x_{jk}$  ( $j = 1, \dots, m; k = 1, \dots, n$ ). The  $x_{jk}$  are realizations of random variables  $X_j$ , varying between specimens and repeated measurements at the same specimen. As in the case of 2 methods we assume that each random variable  $X_j$  can be decomposed in 2 independent random variables  $X_j^*$  and  $E_j$ :

$$X_j = X_j^* + E_j$$

where  $X_j^*$  represents the "expected value" and  $E_j$  the "error term" of the method number  $j$ .

The expected value reflects the random variation of the analyte between different specimens, if there are no errors in the measurement methods. The error term reflects these measurement errors, which vary randomly within repeated measurements of the same specimen.

We further assume that the expected values  $X_1^*, \dots, X_m^*$  have a  $m$ -dimensional, multivariate normal distribution (within the population of all specimens) with the mean vector  $(\mu_1, \dots, \mu_m)$  and the variances  $(\alpha_1^2, \dots, \alpha_m^2)$ . The

covariance matrix of the expected values is determined by the assumption of a *pairwise linear structural relationship*; i.e. we assume, that each pair  $(X_i^*, X_j^*)$  of expected values has the correlation 1 and is therefore with probability 1 linearly interrelated:

$$\frac{X_i^* - \mu_i}{\alpha_i} = \frac{X_j^* - \mu_j}{\alpha_j}$$

This interrelation means that in a  $X_i^* - X_j^*$  coordinate system all possible points  $(X_i^*, X_j^*)$  (formed by the pairs of possible expected values for both analytic methods) lay on a straight line through the mean point  $(\mu_i, \mu_j)$

with a slope  $\beta_{ij} = \frac{\alpha_i}{\alpha_j}$  to the  $X_j^*$ -axis (resp.  $\beta_{ji} = \frac{\alpha_j}{\alpha_i}$  to the  $X_i^*$ -axis).

Considering the expected values of all the  $m$  methods as points in a  $m$ -dimensional euclidean space, the pairwise structural relationship is equivalent to the condition that all possible points of expected values lay on a straight line in the  $m$ -dimensional space. We call this line the *m-dimensional structural relationship line* (see fig. 4).

If we further assume, that the error terms  $E_j$  are independently and normally distributed with mean value 0 and variances  $\lambda_j^2$  — and that the  $E_j$  are independent of the  $X_j^*$  — we have the complete model for the  $m$  measurements  $X_1, \dots, X_m$ . According to this model (which we call model  $\Omega$ ), the  $m$  measurements form a  $m$ -dimensional random vector, which is normally distributed with the mean vector  $(\mu_1, \dots, \mu_m)$  and the variances and covariances:

$$\sigma_j^2 = \alpha_j^2 + \lambda_j^2$$

$$\sigma_{ij} = \alpha_i \cdot \alpha_j$$

$$i, j = 1, \dots, m.$$

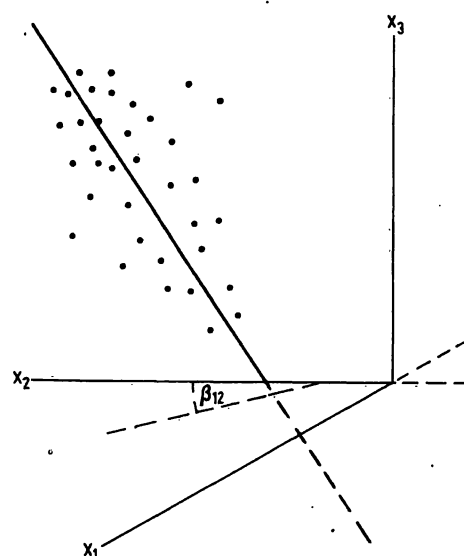


Fig. 4. The structural relationship line between 3 measurement values ( $x_{1k}, x_{2k}, x_{3k}$ ) in the 3-dimensional space.

The structural relationship model  $\Omega$  for  $m$  variables introduces  $3m$  parameters  $\mu_1, \dots, \mu_m, \alpha_1^2, \dots, \alpha_m^2, \lambda_1^2, \dots, \lambda_m^2$ , which relate to  $\frac{(m+3)m}{2}$  parameters (i.e. the means  $\mu_1, \dots, \mu_m$ , the variances  $\sigma_1^2, \dots, \sigma_m^2$  and covariances  $\sigma_{ij}$  ( $i \neq j$ )), as shown by the equation above. The correspondence between the second order moments and the structural parameter  $\alpha_j^2$  and  $\lambda_j^2$  is one-to-one only for the case  $m = 3$ . In this case the structural parameters  $\alpha_j^2$  and  $\lambda_j^2$  can uniquely be calculated by the variances  $\sigma_j^2$  and covariances  $\sigma_{ij}$  without any additional constraints:

$$\alpha_1^2 = \frac{\sigma_{12} \cdot \sigma_{13}}{\sigma_{23}}; \alpha_2^2 = \frac{\sigma_{23} \cdot \sigma_{12}}{\sigma_{13}}; \alpha_3^2 = \frac{\sigma_{13} \cdot \sigma_{23}}{\sigma_{12}}$$

$$\lambda_1^2 = \sigma_1^2 - \alpha_1^2; \lambda_2^2 = \sigma_2^2 - \alpha_2^2; \lambda_3^2 = \sigma_3^2 - \alpha_3^2.$$

The pairwise slopes  $\beta_{ji} = \alpha_j/\alpha_i$ , i.e. the slopes to the  $X_i$ -axis of the projection of the structural relationship line in the  $X_i - X_j$ -plane, are:

$$\beta_{21} = \frac{\alpha_2}{\alpha_1} = \frac{\sigma_{23}}{\sigma_{13}}; \beta_{31} = \frac{\alpha_3}{\alpha_1} = \frac{\sigma_{23}}{\sigma_{12}}; \beta_{32} = \frac{\alpha_3}{\alpha_2} = \frac{\sigma_{13}}{\sigma_{12}}.$$

In the case of  $m > 3$  the relations between the second order moments and structural relationship parameters impose constraints to the variances  $\sigma_j^2$  and covariances  $\sigma_{ij}$  of the variables  $X_1, \dots, X_j$ . This means that these variables cannot vary unrestricted, but only in such a way that the basic relations remain valid. So by the structural relationship not only the variability of the expected values  $X_j^*$ , but also those of the observed values  $X_j$  are restricted if there are more than 3 measurements to compare.

In the case  $m = 3$  the parameters  $\alpha_j^2$  and  $\lambda_j^2$  can be estimated by:

1. estimating  $\sigma_j^2$  and  $\sigma_{ij}$  (with the usual sample variances  $S_j^2$  and covariances  $S_{ij}$ ) and
2. calculating from these estimates the estimates of  $\alpha_j^2$  and  $\lambda_j^2$  (using the unique one-to-one correspondence).

E.g., estimates of the pairwise slopes  $\beta_{ij}$  are:

$$b_{21} = \frac{S_{23}}{S_{13}}; b_{31} = \frac{S_{23}}{S_{12}}; b_{32} = \frac{S_{13}}{S_{12}}$$

where the  $\hat{S}_{ij} = \frac{1}{n} \sum (x_{ik} - \bar{x}_i)(x_{jk} - \bar{x}_j)$  are the (biased) maximum-likelihood estimates for the covariances  $\sigma_{ij}$ .

Approximative values of the sample standard deviations for these estimates can be obtained by using the approximations for functions of sample-moments, as shown in appendix 1 (see also Barnett (12)).

In the case of  $m > 3$  this estimation procedure is not further applicable, as there is no one-to-one correspondence between  $\alpha_j, \lambda_j$  and  $\sigma_{ij}$ . In this case the para-

eters  $\alpha_j^2$  and  $\lambda_j^2$  must be inserted in the likelihood function instead of  $\sigma_j^2$  and  $\sigma_{ij}$ .

This leads to a nonlinear optimization problem, which must be solved numerically. Appropriate computer programs were developed by Feldmann & Klinkers (31). Also the program package LISREL (23) for general structural relationships can be used.

Tests of hypotheses on the parameters can be performed with the likelihood ratio procedure of Neyman and Pearson (see e.g. l.c. (30), (24)). According to this procedure the maximum of the log-likelihood function is calculated without and with the constraints imposed by the hypothesis to be tested. If the hypothesis is valid, the twofold difference between the two log-likelihood values  $L$  (without the constraint  $H_0$ ) and  $L_0$  (with the constraint  $H_0$ ) is asymptotically  $\chi^2$ -distributed with  $df - df_0$  degrees of freedom (where  $df$  is the number of (functionally) independent parameters under the model  $\Omega$  and  $df_0$  under the constrained model  $\Omega \cap H_0$ ).

### 3.2 Example

The application of the structural relationship methods for the comparison of more than 2 methods will be demonstrated by the following example of which the experimental data are taken from l.c. (25):

For the determination of the cholesterol concentration in blood serum  $m = 8$  different methods (PAP, Kageyama, Merckotest, ALDH, ABA, GENT, AA and SMA, abbreviations as used in l.c. (25)) were applied to  $n = 158$  sera. From the results, the mean values  $\bar{x}_j$ , variances  $s_j^2$ , covariances  $s_{ij}^2$  and correlation coefficients  $r_{ij}$  were calculated ( $i, j = 1, \dots, 8$ ). These estimates are shown in table 4.

To estimate the parameters  $\mu, \alpha_j$  and  $\lambda_j$  the likelihood function  $L_0$  for the unconstrained linear functional relationship model was maximized, using the program of l.c. (31). The resulting maximum likelihood ML-estimates  $\hat{\mu}, \hat{\alpha}_j$  and  $\hat{\lambda}_j$  are shown in table 5. The log likelihood without constraints is:  $L = -728.0$ . This model implies  $df = 24$  parameters.

As a measure of precision for each method the coefficient of determination  $R_j^2 = \alpha_j^2/s_j^2$  was calculated (5th column in table 5) and the methods ranked by their precision (6th column in tab. 5). According to this ranking the method PAP has highest precision and the method SMA lowest one. The 8 methods can be considered as equivalent, if their mean values  $\mu_j$  and variance components  $\alpha_j^2$  are identical. Therefore we test the hypothesis:

$$H_0: \mu_1 = \mu_2 = \dots = \mu_8 = \mu$$

and

$$\alpha_1^2 = \alpha_2^2 = \dots = \alpha_8^2 = \alpha^2.$$

This hypothesis can be tested by the likelihood ratio test.

Tab. 4. The complete multivariate sample information of  $m = 8$  methods for the determination of cholesterol concentration with  $n = 158$  samples.  
Coefficients of correlation (upper triangle), variances (main diagonal) and covariances (lower triangle).  
For abbreviations and experimental data see l.c. (14).

Method		Mean of observed measurement $\bar{x}_j$	Correlations – variances – covariances of observed measurements							
$X_j$			$X_1$ PAP	$X_2$ KAG	$X_3$ MERCK	$X_4$ AIDH	$X_5$ ABA	$X_6$ GENT	$X_7$ AA	$X_8$ SMA
$X_1$	PAP	4.784	2.459	0.891	0.985	0.983	0.981	0.974	0.987	0.883
$X_2$	KAGEYAMA	5.167	2.527	3.255	0.880	0.904	0.871	0.903	0.903	0.822
$X_3$	Merckotest	4.562	2.467	2.525	2.543	0.971	0.978	0.980	0.980	0.881
$X_4$	AIDH	4.691	2.387	2.451	2.392	2.387	0.967	0.959	0.979	0.875
$X_5$	ABA	4.957	2.431	2.572	2.460	2.357	2.488	0.974	0.986	0.875
$X_6$	GENT	4.732	2.487	2.554	2.538	2.406	2.495	2.639	0.972	0.882
$X_7$	AA	4.772	2.348	2.467	2.365	2.290	2.354	2.389	2.292	0.889
$X_8$	SMA	5.091	1.894	2.024	1.917	1.845	1.883	1.954	1.836	1.862

Tab. 5. Maximum-likelihood estimates of unconstrained multivariate ( $m = 8$ ) linear functional relationship. The log-likelihood is  $L_0 = -728.0$  with  $df = 24$  parameters.  
For abbreviations and experimental data see l.c. (14).

Method		Mean of expected measurement	Standard deviation of		Precision	Rank
			expected measurement	random error	Coefficient of determination	
$X_j$		$\hat{\mu}_j$	$a_j$	$l_j$	$R_j^2$	
$X_1$	PAP	4.784	1.562	0.1685	0.988	1
$X_2$	KAGEYAMA	5.167	1.624	0.7856	0.810	7
$X_3$	Merckotest	4.562	1.577	0.2357	0.978	3
$X_4$	AIDH	4.691	1.520	0.2756	0.968	5
$X_5$	ABA	4.957	1.560	0.2352	0.978	4
$X_6$	GENT	4.732	1.594	0.3148	0.962	6
$X_7$	AA	4.772	1.504	0.1723	0.987	2
$X_8$	SMA	5.091	1.216	0.6193	0.794	8

Tab. 6. Maximum-likelihood estimates of multivariate linear functional relationship under the constraint of both equal expected means and equal expected standard deviations (hypothesis of equal accuracies). The log-likelihood is  $L_0 = -877.9$  with  $df = 10$  parameters.  
For abbreviations and experimental data see l.c. (14).

Method		Mean of expected measurement	Standard deviation of		Precision	Rank
			expected measurement	random error	Coefficient of determination	
$X_j$		$\hat{\mu}_j$	$a_j$	$l_j$	$R_j^2$	
$X_1$	PAP	4.773	1.539	0.1662	0.988	1
$X_2$	KAGEYAMA	4.773	1.539	0.8852	0.751	8
$X_3$	Merckotest	4.773	1.539	0.3324	0.955	5
$X_4$	AIDH	4.773	1.539	0.2815	0.968	3
$X_5$	ABA	4.773	1.539	0.3117	0.961	4
$X_6$	GENT	4.773	1.539	0.3324	0.955	6
$X_7$	AA	4.773	1.539	0.1732	0.987	2
$X_8$	SMA	4.773	1.539	0.7686	0.800	7

We have to calculate the estimates for  $\mu$ ,  $\alpha^2$  and  $\lambda_j^2$  ( $j = 1, \dots, 8$ ) and the log likelihood  $L_0$  under the constraint  $H_0$ . The results are shown in table 6. The log likelihood under the constraint  $H_0$  is:  $L_0 = -877.9$  with  $df = 10$  parameters. Therefore we get as test statistics:

$$X^2 = 2(L - L_0) = 299.8.$$

This statistic is asymptotic  $\chi^2$ -distributed with  $df - df_0 = 14$  degrees of freedom. For an error probability  $P < 0.001$  we get  $\chi_0^2 = 36.1$ . As  $X^2 > \chi_0^2$  the hypothesis  $H_0$  must be rejected on a level  $< 0.1\%$ . The 8 methods cannot be considered as equivalent. They differ very significantly in their location and slope parameters. These differences can be interpreted as differences in the "accuracy" between the 8 methods.

In table 7 the results of the likelihood ratio tests for several hypotheses are shown. This table demonstrates that the 8 methods differ highly significant not only in accuracy but also in precision. The coefficients of determination  $R_j^2$  differ very significantly between PAP and *Kageyama*, but not between PAP and AA. The methods Mercktest, ALDH and ABA differ significantly in their location parameters, but not in their precision and slopes ( $R_j^2$ ,  $\alpha_j^2$ ). It should be observed that the projection of the unique structural relationship line for all methods to the plane of the PAP- and *Kageyama*-methods lies between the principal component and standardized principal component lines derived by the comparison of these two methods.

## Appendix 1

### a) Approximate variances for functions of sample moments

Let  $(X, Y)$  be a bivariate, normally distributed random vector with mean values  $\mu_x, \mu_y$ , variances  $\sigma_x^2, \sigma_y^2$  and

covariance  $\sigma_{xy}$ ;  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$  should be a sample of  $n$  independent realizations of  $(X, Y)$ .

The sample moments are defined as:

$$\bar{x} = \frac{1}{n} \sum_i x_i \quad \bar{y} = \frac{1}{n} \sum_i y_i$$

$$S_x^2 = \frac{1}{n} \sum_i (x_i - \bar{x})^2; \quad S_y^2 = \frac{1}{n} \sum_i (y_i - \bar{y})^2;$$

$$S_{xy} = \frac{1}{n} \sum_i (x_i - \bar{x})(y_i - \bar{y})$$

It is well known (see e.g. l.c. (8) chapter 27 and 28), that these sample moments are asymptotically normally distributed; their mean values equal the corresponding parameters up to an error of order  $O(1/n)$  and the variances and covariances of the second order moments are:

$$\text{var}(S_x^2) = \frac{2}{n} \sigma_x^4 + O(1/n^2)$$

$$\text{var}(S_y^2) = \frac{2}{n} \sigma_y^4 + O(1/n^2)$$

$$\text{var}(S_{xy}) = \frac{1}{n} (\sigma_{xy}^2 + \sigma_x^2 \sigma_y^2) + O(1/n^2)$$

$$\text{cov}(S_x^2, S_y^2) = \frac{2}{n} \sigma_{xy}^2 + O(1/n^2)$$

$$\text{cov}(S_x^2, S_{xy}) = \frac{2}{n} \sigma_{xy} \cdot \sigma_x^2 + O(1/n^2)$$

$$\text{cov}(S_y^2, S_{xy}) = \frac{2}{n} \sigma_{xy} \cdot \sigma_y^2 + O(1/n^2)$$

### Theorem of Cramér

Let  $H(S_x^2, S_y^2, S_{xy})$  be a function of the second order sample moments, continuous and continuous differentiable in  $(\sigma_x^2, \sigma_y^2, \sigma_{xy})$  with  $H < Cn^p$  for all sample values  $(x_i, y_i)$  and  $C \geq 0, p \geq 0$ .

Tab. 7. The likelihood-ratio-test of several hypotheses. Error of first kind:  $-P \geq 0.05$ ,  $+P < 0.05$ ,  $++P < 0.01$ ,  $+++P < 0.001$ . The suffixes correspond to the 8 methods from which the experimental data (l.c. (14)) are taken as in the previous tables.

Hypothesis	Log-likelihood	Parameter	Likelihood-ratio-test		
Unconstrained model	- 727.996	24	$\chi^2$	df	P
$\mu_1 = \mu_2 = \dots = \mu_8$ and $\alpha_1 = \alpha_2 = \dots = \alpha_8$	- 877.9	10	299.8	14	+++
$\mu_1 = \mu_2 = \dots = \mu_8$	- 840.8	17	225.6	7	+++
$\alpha_1 = \alpha_2 = \dots = \alpha_8$	- 757.8	17	59.6	7	+++
$\lambda_1 = \lambda_2 = \dots = \lambda_8$	-1035.7	17	615.4	7	+++
$\rho_1 = \rho_2 = \dots = \rho_8$	-1027.7	17	599.4	7	+++
$\rho_1 = \rho_2$	- 827.2	23	194.4	1	+++
$\rho_1 = \rho_7$	- 728.014	23	0.036	1	-
$\mu_1 = \mu_7$ and $\alpha_1 = \alpha_7$	- 783.9	22	111.8	2	+++
$\rho_3 = \rho_4 = \rho_5$	- 730.5	22	5.01	2	-
$\alpha_3 = \alpha_4 = \alpha_5$ and $\rho_3 = \rho_4 = \rho_5$	- 731.8	20	7.61	4	-
$\mu_3 = \mu_4 = \mu_5$ and $\alpha_3 = \alpha_4 = \alpha_5$	- 811.7	20	167.4	4	+++

Then the following relations are valid:

$$E(H) = H_0 + 0(1/n)$$

$$\begin{aligned} \text{var}(H) = & H_x^2 \text{var}(S_x^2) + H_y^2 \text{var}(S_y^2) + H_{xy}^2 \text{var}(S_{xy}) + \\ & + 2H_x \cdot H_y \text{cov}(S_x^2, S_y^2) + \\ & + 2H_x \cdot H_{xy} \text{cov}(S_x^2, S_{xy}) + \\ & + 2H_y \cdot H_{xy} \text{cov}(S_y^2, S_{xy}) + 0(n^{-3/2}) \end{aligned}$$

where:

$$H_0 = H(\sigma_x^2, \sigma_y^2, \sigma_{xy})$$

$$H_x = \frac{\partial}{\partial S_x^2} H(\sigma_x^2, \sigma_y^2, \sigma_{xy});$$

$$H_y = \frac{\partial}{\partial S_y^2} H(\sigma_x^2, \sigma_y^2, \sigma_{xy})$$

$$H_{xy} = \frac{\partial}{\partial S_{xy}} H(\sigma_x^2, \sigma_y^2, \sigma_{xy})$$

The proof of this theorem is given by *Cramér* ((8), chapter 27.7).

#### b) Application to the slope-estimates of the structural relationship line

With the constraint:  $\Theta^2 = \sigma_y^2 / \sigma_x^2$  (= a given, fix value) the ML-estimate for the slope  $\beta_1$  of the structural relationship line is:

$$\begin{aligned} b_1(\Theta) = & \frac{1}{2S_{xy}} (S_y^2 - \Theta^2 S_x^2 + \\ & + \sqrt{(S_y^2 - \Theta^2 S_x^2)^2 + 4\Theta^2 S_{xy}^2}) \end{aligned}$$

Considering this estimate as a function  $H$  of the second order moments we have:

$$H_x = -\frac{\Theta^2}{2\sigma_{xy}} \cdot K$$

$$H_y = \frac{1}{2\sigma_{xy}} \cdot K$$

$$H_{xy} = -\frac{(\sigma_y^2 - \Theta^2 \sigma_x^2)}{2\sigma_{xy}^2} \cdot K$$

with

$$K = 1 + (\sigma_y^2 - \Theta^2 \sigma_x^2) / \sqrt{(\sigma_y^2 - \Theta^2 \sigma_x^2)^2 + 4\Theta^2 \sigma_{xy}^2}$$

So we get from the theorem of *Cramér*:

$$\begin{aligned} \text{var}(b_1(\Theta)) = & \frac{K^2}{4\sigma_{xy}^2 \cdot n} (\sigma_y^2 - \Theta^2 \sigma_x^2)^2 \frac{\sigma_x^2 \sigma_y^2 - 3\sigma_{xy}^2}{\sigma_{xy}^2} + \\ & + 2(\sigma_y^4 - 2\Theta^2 \sigma_x^2 \sigma_{xy}^2 + \Theta^4 \sigma_x^4) + 0(n^{-3/2}) \end{aligned}$$

From this formula approximate expressions for the variance of the slope with special constraints (e.g. princi-

pal component constraint  $\Theta = 1$  or regression constraint  $\Theta = 0$ ) can be derived. So we get for the regression of  $x$  on  $y$  with the constraint  $\Theta = 0$ :

$$b_1(0) = S_y^2 / S_{xy}$$

$$\begin{aligned} \text{var}(b_1(0)) = & \frac{1}{n} \cdot \frac{\sigma_y^4}{\sigma_{xy}^2} \left( \frac{\sigma_x^2 \cdot \sigma_y^2}{\sigma_{xy}^2} - 1 \right) + 0(n^{-3/2}) = \\ = & \frac{1}{n} \cdot \frac{\sigma_y^4}{\sigma_{xy}^2} \frac{1 - \rho^2}{\rho^2} + 0(n^{-3/2}) \end{aligned}$$

The approximate variance for  $b_1$  in the regression of  $y$  on  $x$  with the constraint  $\Theta \rightarrow \infty$  can be derived more easily by applying the theorem of *Cramér* to the estimate:

$$b_1(\infty) = \frac{S_{xy}}{S_x^2} = H(S_x^2, S_{xy})$$

So we get:

$$\begin{aligned} H_x = & -\frac{\sigma_{xy}}{\sigma_x^4}; \quad H_{xy} = \frac{1}{\sigma_x^2} \\ \text{var}(b_1(\infty)) = & \frac{2}{n} \frac{\sigma_{xy}^2}{\sigma_x^4} + \frac{1}{n} \frac{\sigma_{xy}^2}{\sigma_x^4} + \frac{1}{n} \frac{\sigma_y^2}{\sigma_x^2} - \\ & - \frac{4}{n} \frac{\sigma_{xy}^2}{\sigma_x^4} + 0(n^{-3/2}) = \\ = & \frac{1}{n\sigma_x^2} (\sigma_y^2 - \frac{\sigma_{xy}^2}{\sigma_x^2}) + 0(n^{-3/2}) = \\ = & \frac{1}{n} \cdot \frac{\sigma_y^2(1 - \rho^2)}{\sigma_x^2} + 0(n^{-3/2}). \end{aligned}$$

In the case of the standardized principal component the constraint  $\Theta$  is not fixed but depends on the slope  $\beta_1$ , which is a function of  $\sigma_x^2$  and  $\sigma_y^2$ . The easiest way to get an approximate expression for the variance is to apply the theorem of *Cramér* directly to the estimate:

$$b_1 = \frac{S_y}{S_x} = (S_y^2)^{1/2} \cdot (S_x^2)^{-1/2} = H(S_x^2, S_y^2)$$

$$H_x = -\frac{1}{2} \frac{\sigma_y}{\sigma_x^3}; \quad H_y = \frac{1}{2} \frac{1}{\sigma_y \cdot \sigma_x}$$

$$\begin{aligned} \text{var}(b_1) = & \frac{1}{2n} \frac{\sigma_y^2}{\sigma_x^2} + \frac{1}{2n} \frac{\sigma_y^2}{\sigma_x^2} - \frac{1}{n} \frac{\sigma_{xy}^2}{\sigma_x^4} + 0(n^{-3/2}) = \\ = & \frac{1}{n} \frac{\sigma_y^2}{\sigma_x^2} \left( 1 - \frac{\sigma_{xy}^2}{\sigma_x^2 \sigma_y^2} \right) + 0(n^{-3/2}) = \\ = & \frac{1}{n} \beta_1^2 (1 - \rho^2) + 0(n^{-3/2}) \end{aligned}$$



## References

1. Averdunk, R. & Borner, K. (1970), *Z. Klin. Chem. Klin. Biochem.* **8**, 263–268.
2. Cornbleet, P. & Gochman, N. (1979), *J. Clin. Chem.* **25**, 432–438.
3. Grubbs, F. E. (1948), *J. Amer. Statist. Assoc.* **43**, 243–264.
4. Haeckel, R. (1975), *Qualitätssicherung im medizinischen Laboratorium*. Deutscher Ärzteverlag, Köln.
5. Theobald, C. M. & Mallison, J. R. (1978), *Biometrics* **44**, 39–45.
6. Wakkers, P. J. M., Mellendoorn, H. B. A., Op de Weega, G. J. & Heerspink, W. (1975), *Clin. Chim. Acta* **64**, 173–184.
7. Deming, W. E. (1943), *Statistical adjustment of data*. Wiley, New York.
8. Crámer, H. (1958), *Mathematical Methods of Statistics*. Princeton, University Press. 8th printing.
9. Pearson, K. (1901), *Phil. Mag. Ser. 6*, 559–572.
10. Anderson, T. W. (1958), *An introduction of multivariate statistical analysis*. Wiley, New York.
11. Barnett, V. D. (1967), *Biometrika* **54**, 670–671.
12. Barnett, V. D. (1969), *Biometrics* **25**, 129–142.
13. Geary, R. C. (1942), *Proc. Roy. Irish Acad. Ser. 47*, 63–76.
14. Madansky, A. (1959), *Journ. Amer. Stat. Ass.* **54**, 173–205.
15. Moran, P. A. P. (1971), *J. Mult. Anal.* **1**, 232–255.
16. Neyman, J. (1951), *Ann. Math. Stat.* **22**, 496–512.
17. Reiersøl, O. (1945), *Ark. Mat. Astron. Fys.* **32**, 1–119.
18. Wolfowitz, J. (1957), *Ann. Math. Stat.* **28**, 75–88.
19. Cox, N. R. (1976), *Biometrika* **63**, 231–237.
20. Gauss, C. F. (1821), *Theorie der den kleinsten Fehlern unterworfenen Combination der Beobachtungen*. Abhandlungen zur Methode der kleinsten Quadrate von C. F. Gauss. Physica Verlag, Würzburg 1964.
21. Hotelling, H. (1939), *Psychometrika* **1**, 27–35.
22. Lawley, D. N. & Maxwell, A. E. (1973), *Biometrika* **60**, 331–338.
23. LISREL IV, *Analysis of Linear Structural Relationships by the Method of Maximum Likelihood*. Edited by: Jöreskog, K. G. & Sörbom, D. National Educational Resources. Inc., Chicago 1978.
24. Rao, C. R. (1965), *Linear statistical inference and its applications*. Wiley, New York.
25. Haeckel, R., Sonntag, O., Külpmann, W. R., Feldmann, U. & Schneider, B. (1979), *J. Clin. Chem. Clin. Biochem.* **17**, 553–563.
26. Wald, A. (1940), *Ann. Math. Stat.* **11**, 284–300.
27. Neymann, J. & Scott, E. L. (1951), *Ann. Math. Stat.* **22**, 352–361. Correction: *Ann. Math. Stat.* **23** (1952), 115.
28. Wolfowitz, J. (1952), *Skand. Aktuarietidskr.*, 132–151.
29. Kiefer, J. (1964), *Ann. Math. Stat.* **35**, 1371–1380.
30. Morgan, W. A. (1939), *Biometrika* **31**, 13–19.
31. Feldmann, U. & Klinkers, H. (1979), *A FORTRAN IV-program for solving linear relationship*. In preparation.

Prof. Dr. B. Schneider  
Medizinische Hochschule Hannover  
Abteilung für Biometrie  
Karl-Wiechert-Allee 9  
D-3000 Hannover 61

